Supporting Information

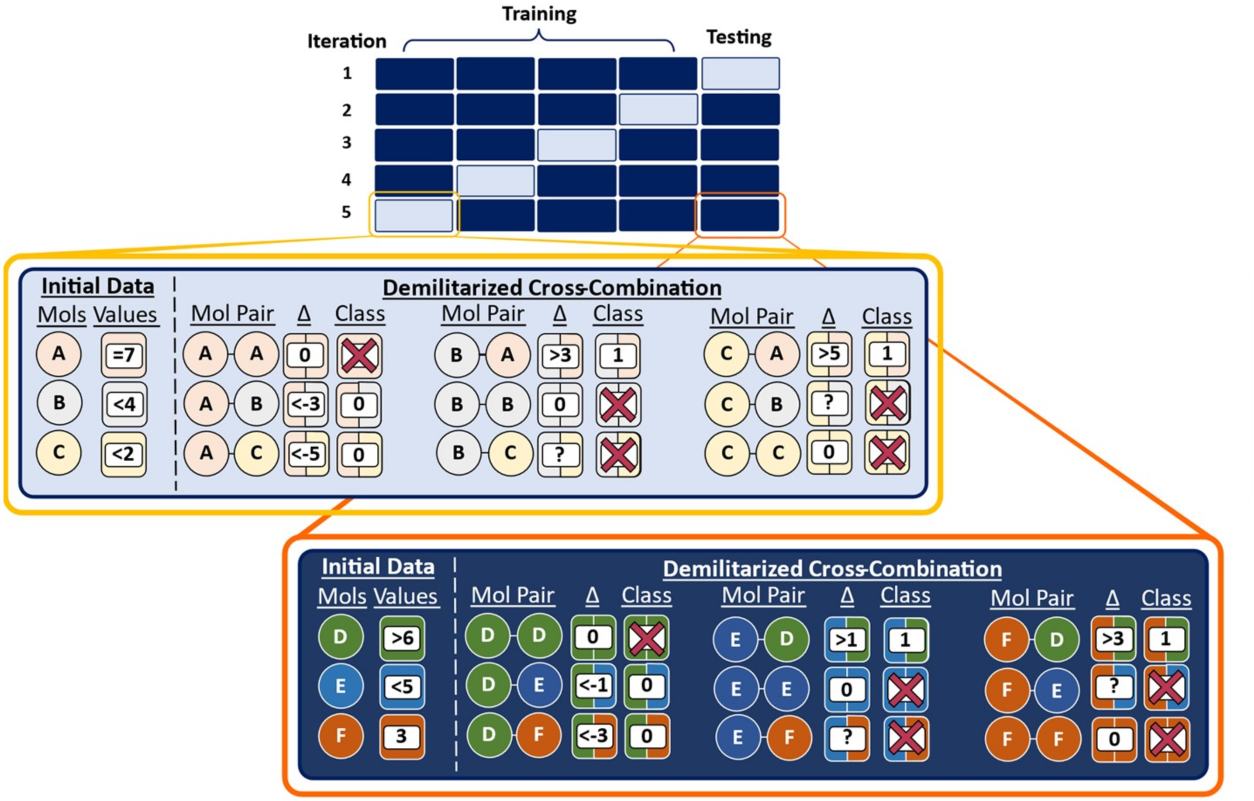
Leveraging bounded datapoints to classify molecular potency improvements

Zachary Fralish[1], Paul Skaluba[1], & Daniel Reker[1]*

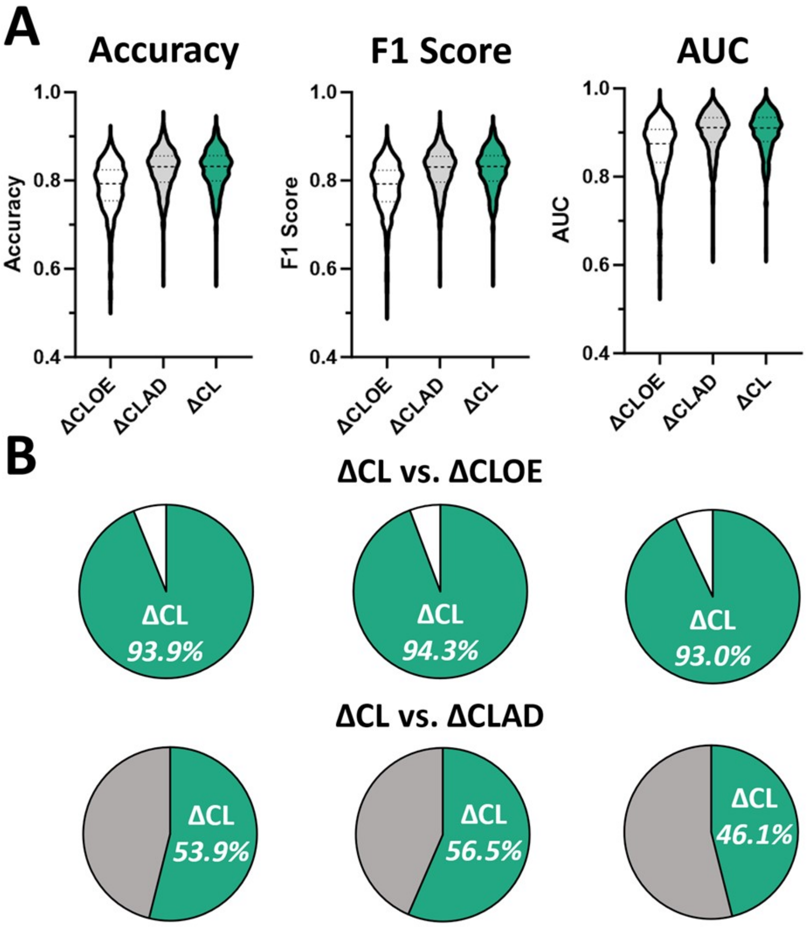
[1] *Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA*

* Corresponding Author: Daniel Reker, daniel.reker@duke.edu

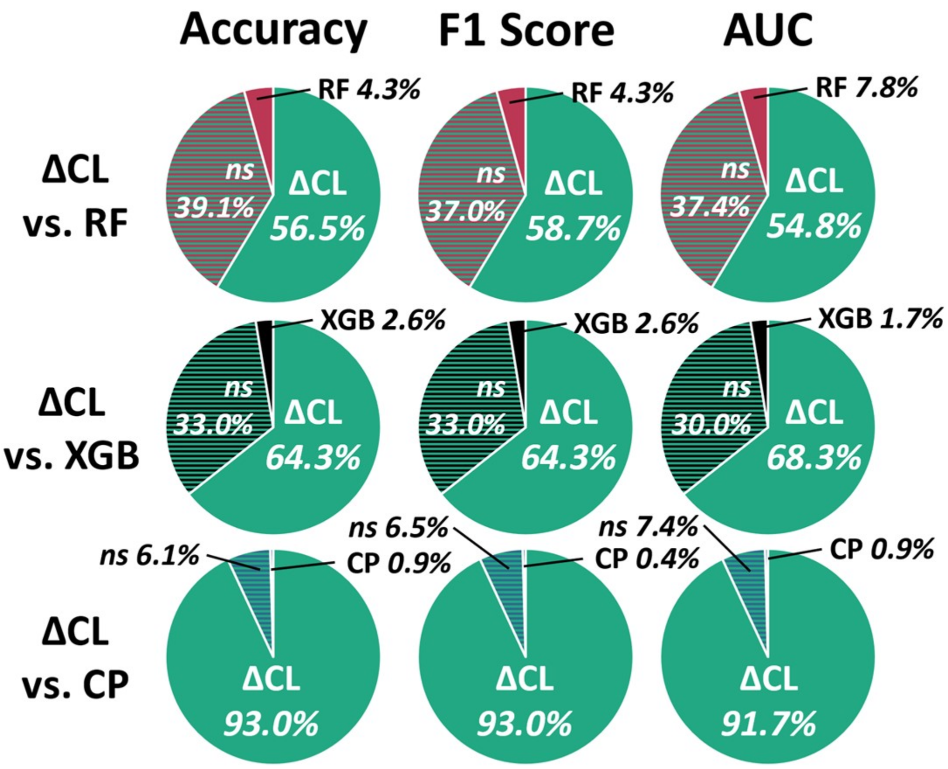
Supplementary Figures



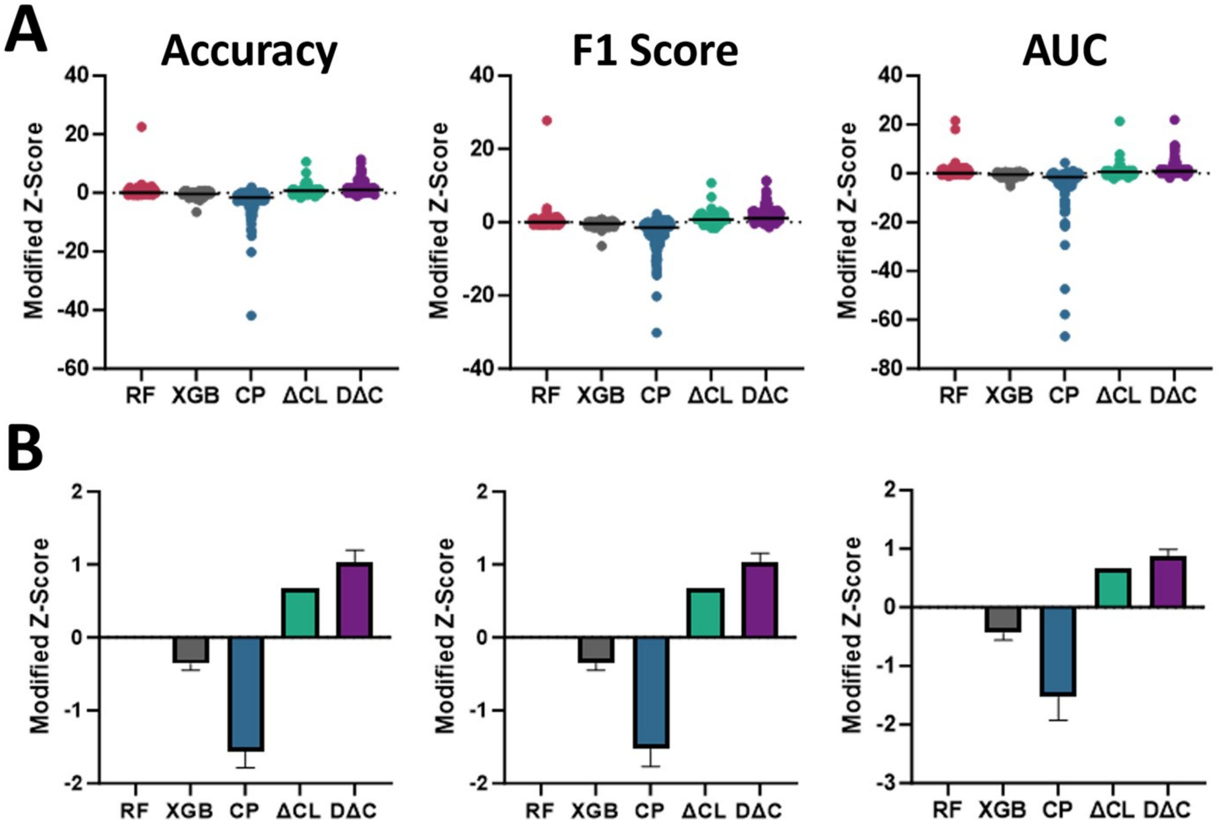
Supplementary Figure 1: Cross-Validation Scheme for DeltaClassifiers. Datapoints undergo cross-merging to generate pairs following cross-validation splits to circumvent data leakage risks. As such, each molecule from the original dataset only occurs in molecule pairs within the training or testing data splits, but never both. Additionally, if it is unknown if the property is improved (e.g., both molecules' properties are denoted as '>') or the difference is less than 0.1 pIC_{50}, the pair is removed.



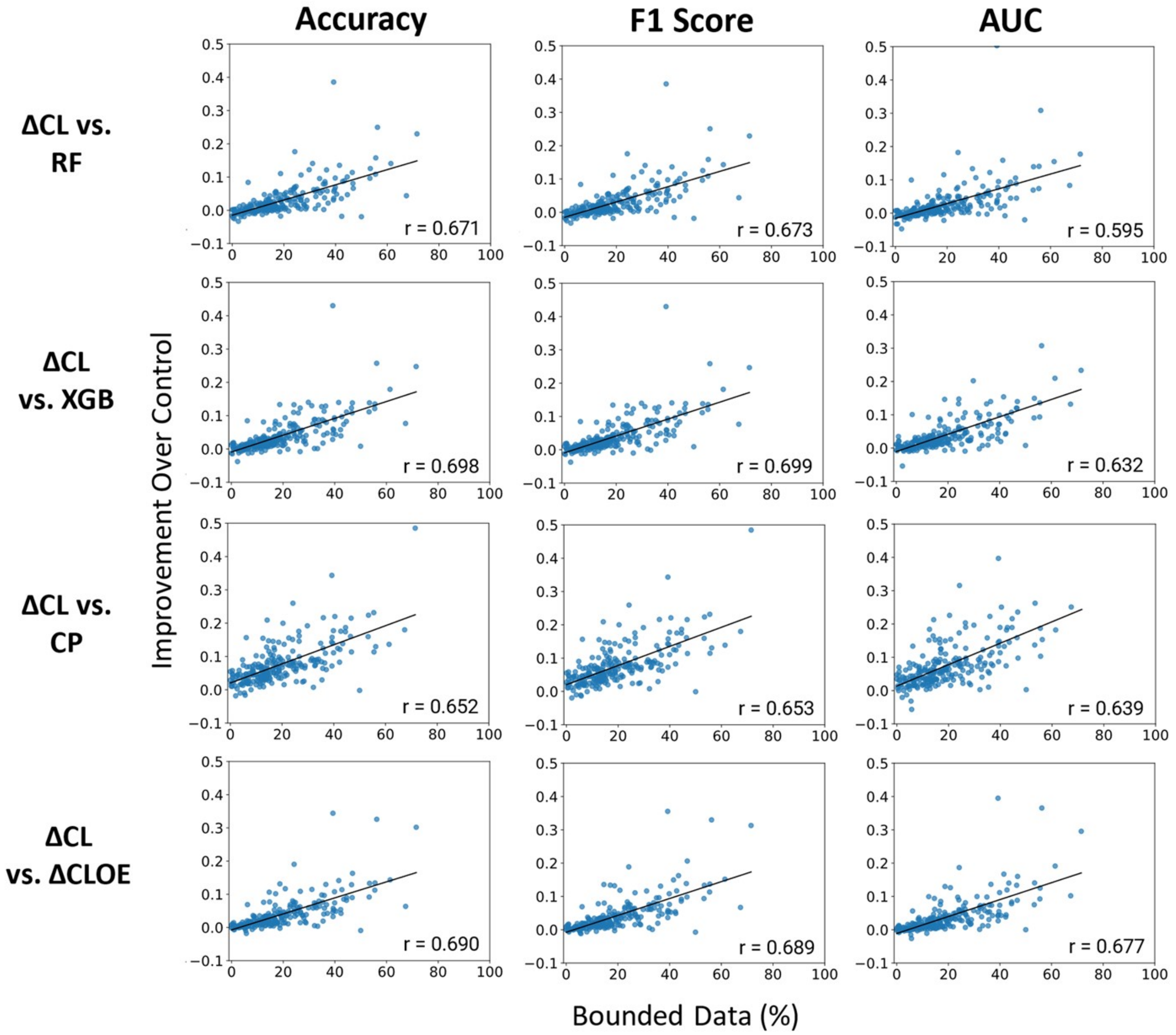
Supplementary Figure 2: Tree-based DeltaClassifier Performance Following Training with Only Exact Values (ΔCLOE), All Data (ΔCLAD), and Demilitarized Data (ΔCL) Tested on Demilitarized Data. (A) Violin plots of model performance following 1x10 cross-validation for 230 ChEMBL datasets in terms of accuracy, F1 score, and AUC. **(B)** Pie charts showing percentage of datasets ΔCL outcompeted ΔCLOE and ΔCLAD.



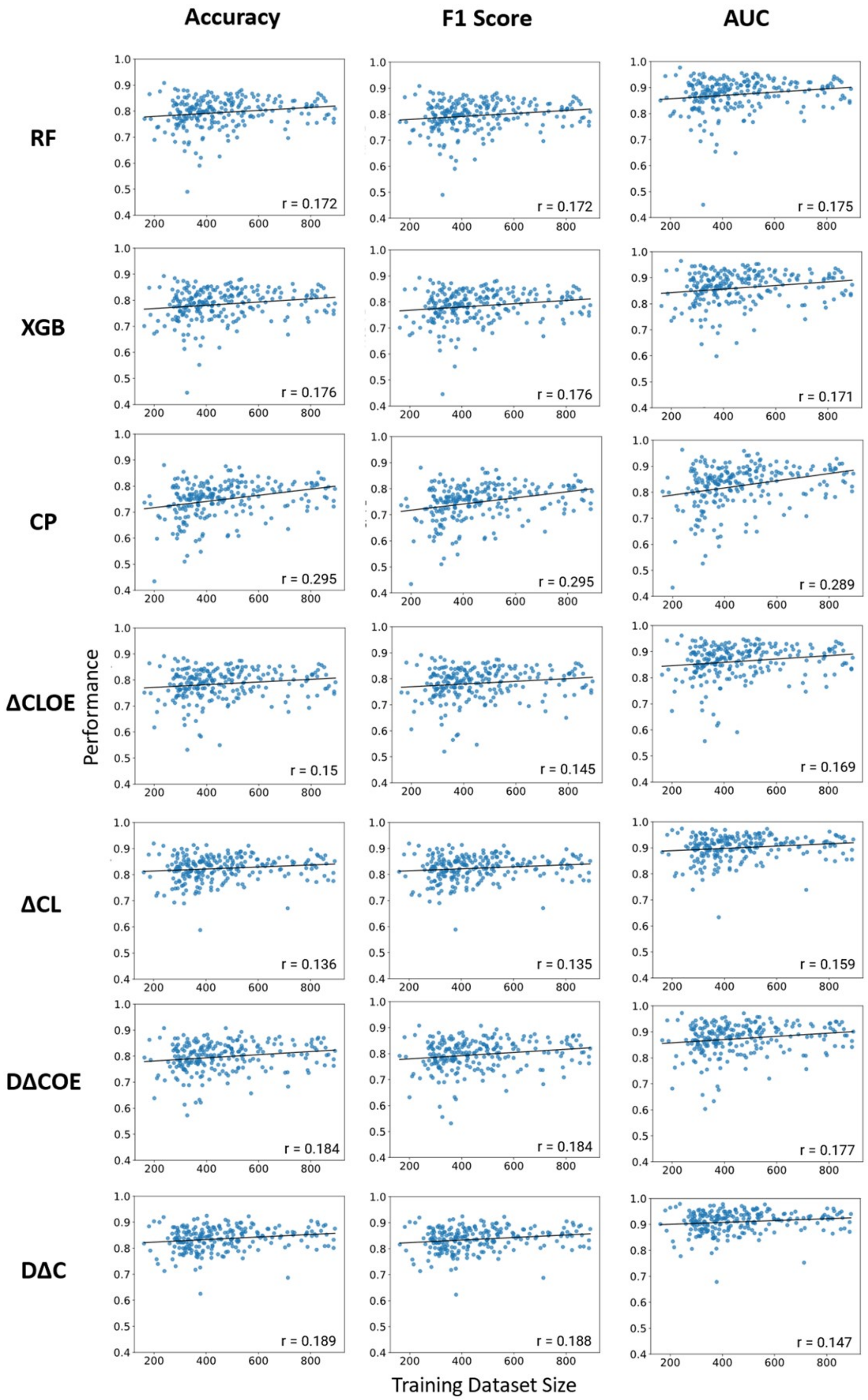
Supplementary Figure 3: Tree-based DeltaClassifier Performance Compared with Traditional Models. Pie charts showing percentage of datasets the tree-based DeltaClassifier (ΔCL) outcompeted (green), exhibited a non-significant difference (gradient), or underperformed Random Forest (RF, red), XGBoost (XGB, black), and Chemprop (CP, blue) during 3x10-fold cross-validation. Statistical significance from paired t-test for three repeats ($p < 0.05$). Note that the DeltaClassifierLite is based on XGBoost. The difference is that the DeltaClassifiers run these algorithms in classification mode after creating paired training data while the standard implementations, including Random Forest and Chemprop, run in regression mode.



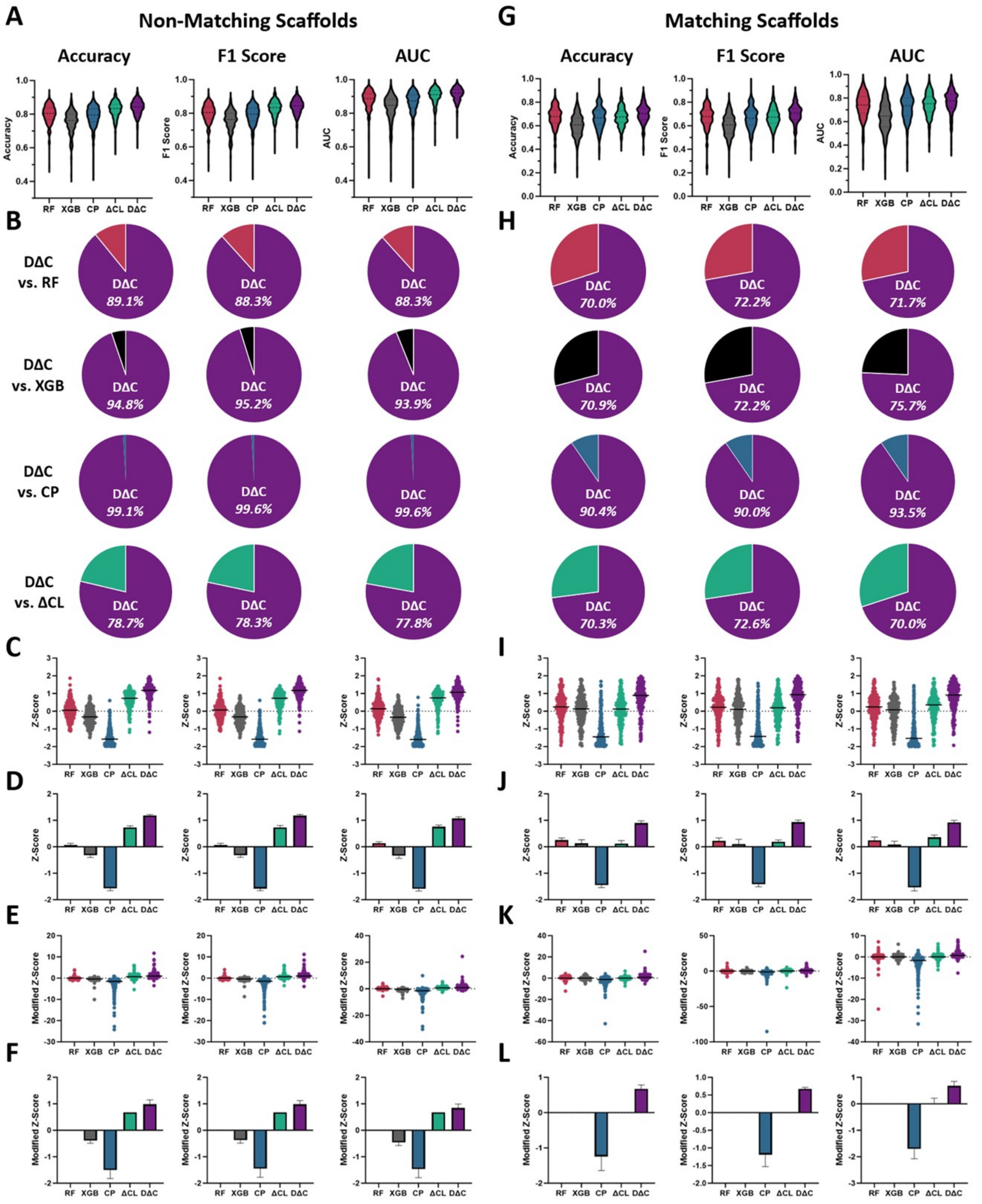
Supplementary Figure 4: Modified Z-Score Calculations. (A) Average modified Z-scores for model (Random Forest (RF), XGBoost (XGB), Chemprop (CP), tree-based DeltaClassifer (ΔCL), and DeltaClassifer (DΔC)) performance following 3x10 cross-validation for 230 ChEMBL datasets in terms of accuracy, F1 score, and AUC. **(B)** Median and 95% confidence interval of average modified z-scores. Note that the DeepDeltaClassifer uses the neural network implementation of Chemprop and the DeltaClassifierLite is based on XGBoost. The difference is that the DeltaClassifiers run these algorithms in classification mode after creating paired training data while the standard implementations, including Random Forest, run in regression mode.



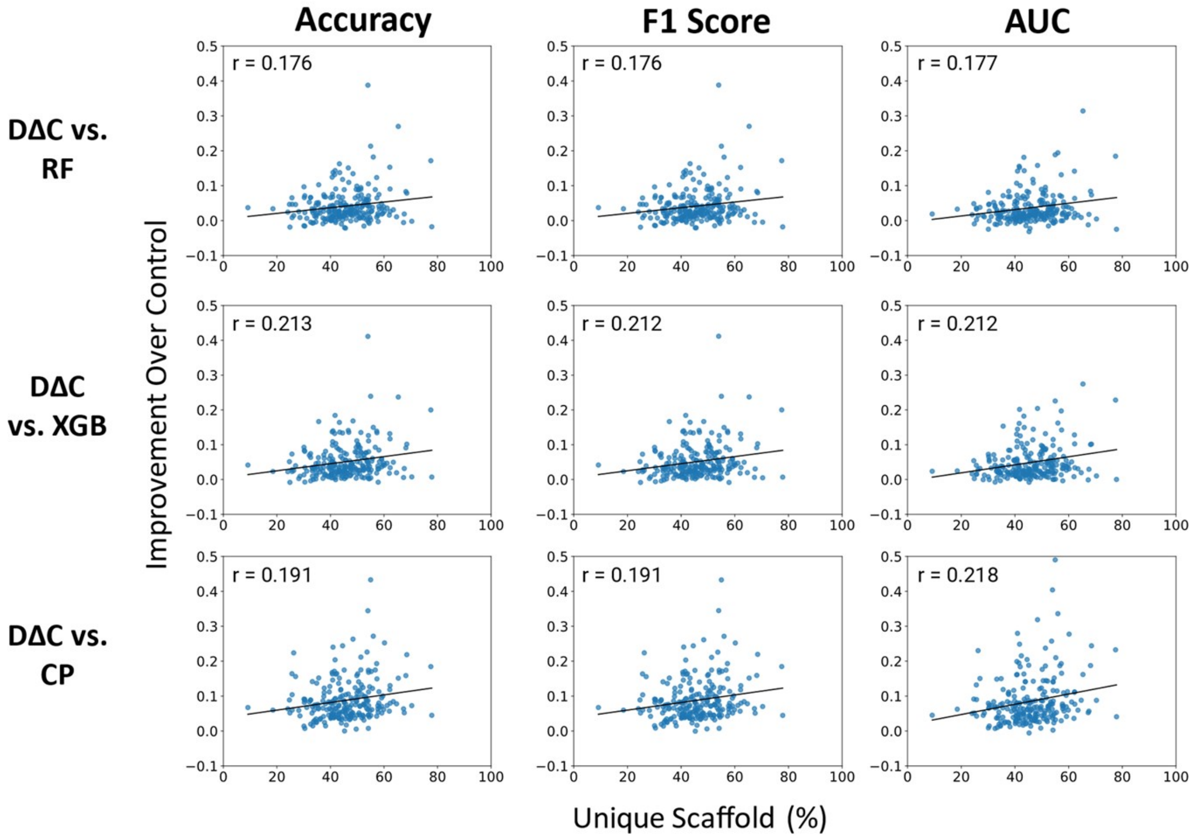
Supplementary Figure 5: Percent of Bounded Data Correlates with ΔCL Improvement Over Traditional Models. Scatterplots showing correlation and Pearson's r values of tree-based DeltaClassifer (ΔCL) performance improvement over Random Forest (RF), XGBoost (XGB) Chemprop (CP), and tree-based DeltaClassifer trained only on exact values (ΔCLOE) following 1x10 cross-validation for 230 ChEMBL datasets with the percent of bounded data within each dataset in terms of accuracy, F1 score, and AUC. Note that the DeepDeltaClassifer uses the neural network implementation of Chemprop and the DeltaClassifierLite is based on XGBoost. The difference is that the DeltaClassifiers run these algorithms in classification mode after creating paired training data while the standard implementations, including Random Forest, run in regression mode.


Accuracy
F1 Score
AUC
RF
XGB
CP
ΔCLOE
ΔCL
DΔCOE
DΔC
Performance
Training Dataset Size
1.0
0.9
0.8
0.7
0.6
0.5
0.4
200
400
600
800
r = 0.172
r = 0.172
r = 0.175
r = 0.176
r = 0.176
r = 0.171
r = 0.295
r = 0.295
r = 0.289
r = 0.15
r = 0.145
r = 0.169
r = 0.136
r = 0.135
r = 0.159
r = 0.184
r = 0.184
r = 0.177
r = 0.189
r = 0.188
r = 0.147

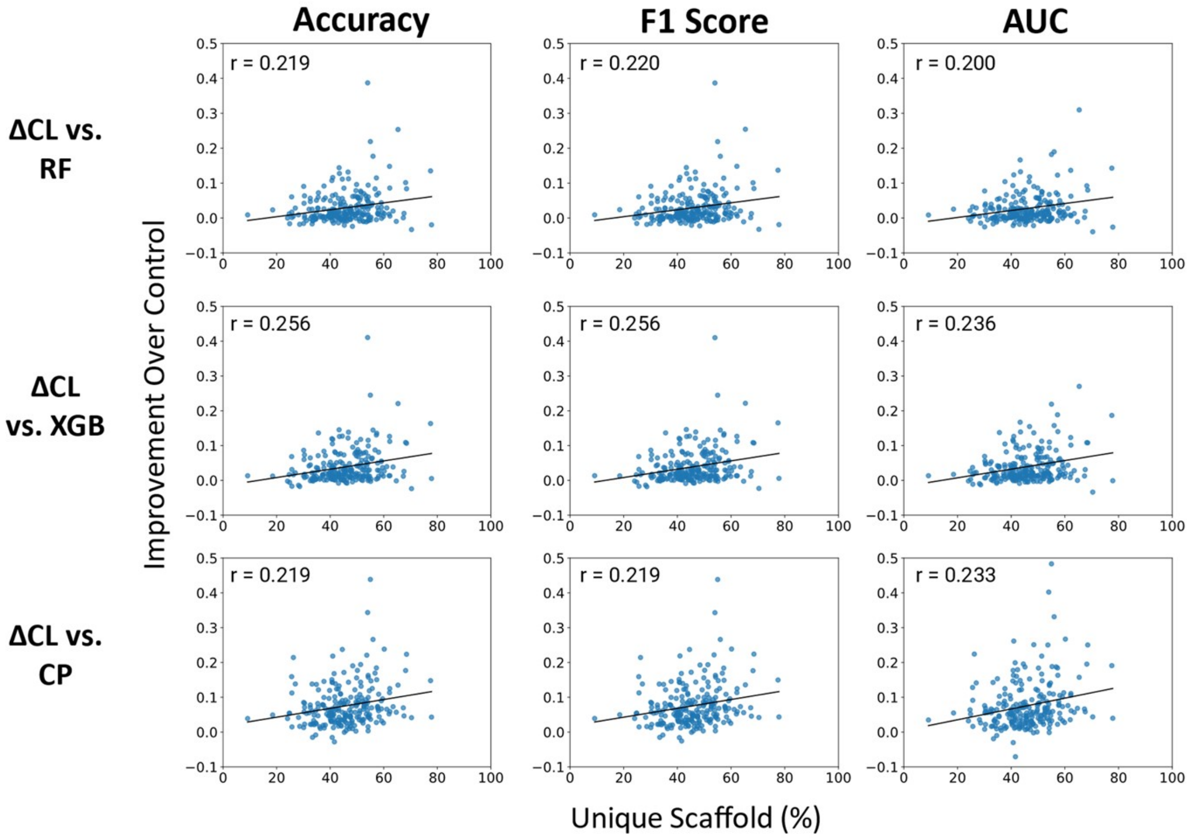
Supplementary Figure 6: Limited Correlation of Dataset Size with Model Performance. Scatterplots showing correlation and Pearson's r values of model performance following 1x10 cross-validation for 230 ChEMBL datasets with dataset size in terms of accuracy, F1 score, and AUC with dataset size for Random Forest (RF), XGBoost (XGB), Chemprop (CP), tree-based DeltaClassifer trained on only exact data (ΔCLOE), tree-based DeltaClassifer (ΔCL), deep DeltaClassifer trained on only exact data (DΔCOE), and deep DeltaClassifer (DΔC). Note that the DeepDeltaClassifer uses the neural network implementation of Chemprop and the DeltaClassifierLite is based on XGBoost. The difference is that the DeltaClassifiers run these algorithms in classification mode after creating paired training data while the standard implementations, including Random Forest, run in regression mode.


A
Non-Matching Scaffolds
Accuracy
F1 Score
AUC
G
Matching Scaffolds
Accuracy
F1 Score
AUC
RF
XGB
CP
ΔCL
DΔC
B
DΔC vs. RF
DΔC 89.1%
DΔC 88.3%
DΔC 88.3%
DΔC vs. XGB
DΔC 94.8%
DΔC 95.2%
DΔC 93.9%
DΔC vs. CP
DΔC 99.1%
DΔC 99.6%
DΔC 99.6%
DΔC vs. ΔCL
DΔC 78.7%
DΔC 78.3%
DΔC 77.8%
H
DΔC 70.0%
DΔC 72.2%
DΔC 71.7%
DΔC 70.9%
DΔC 72.2%
DΔC 75.7%
DΔC 90.4%
DΔC 90.0%
DΔC 93.5%
DΔC 70.3%
DΔC 72.6%
DΔC 70.0%
C
I
Z-Score
D
J
Z-Score
E
K
Modified Z-Score
F
L
Modified Z-Score

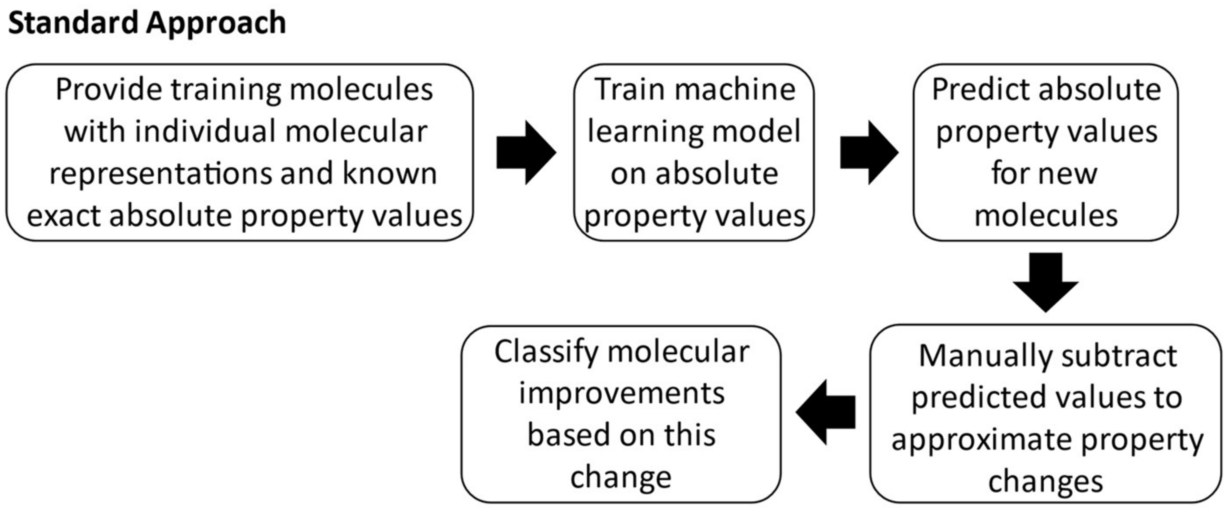
Supplementary Figure 7: Comparison of DeltaClassifiers with Traditional Methods Across Matching or Non-Matching Scaffolds. **(A)** Violin plots of model performance following 1x10 cross-validation for non-matching scaffold pairs for 230 ChEMBL datasets in terms of accuracy, F1-score, and ROCAUC. **(B)** Pie charts showing percentage of datasets our DeepDeltaClassifer (DΔC) outcompeted Random Forest (RF), XGBoost (XGB), Chemprop (CP), and DeltaClassiferLite (ΔCL) in terms of accuracy, F1-score, and ROCAUC for non-matching scaffold pairs. **(C)** Z-scores for model performance in terms of accuracy, F1 score, and ROCAUC for non-matching scaffolds. **(D)** Median and 95% confidence interval of z-scores for non-matching scaffold pairs. **(E)** Modified Z-scores for model performance for non-matching scaffold pairs following 1x10 cross-validation for 230 ChEMBL datasets in terms of accuracy, F1 score, and AUC. **(F)** Median and 95% confidence interval of modified z-scores for non-matching scaffold pairs. **(G)** Violin plots of model performance following 1x10 cross-validation for matching scaffold pairs for 230 ChEMBL datasets in terms of accuracy, F1-score, and ROCAUC. **(H)** Pie charts showing percentage of datasets our DΔC outcompeted RF, XGB, CP, and ΔCL in terms of accuracy, F1-score, and ROCAUC for matching scaffold pairs. **(I)** Z-scores for model performance in terms of accuracy, F1 score, and ROCAUC for matching scaffold pairs. **(J)** Median and 95% confidence interval of z-scores for matching scaffold pairs. **(K)** Modified Z-scores for model performance for matching scaffold pairs following 1x10 cross-validation for 230 ChEMBL datasets in terms of accuracy, F1 score, and AUC. **(L)** Median and 95% confidence interval of modified z-scores for matching scaffold pairs. Note that the DeepDeltaClassifer uses the neural network implementation of Chemprop and the DeltaClassifierLite is based on XGBoost. The difference is that the DeltaClassifiers run these algorithms in classification mode after creating paired training data while the standard implementations, including Random Forest, run in regression mode.



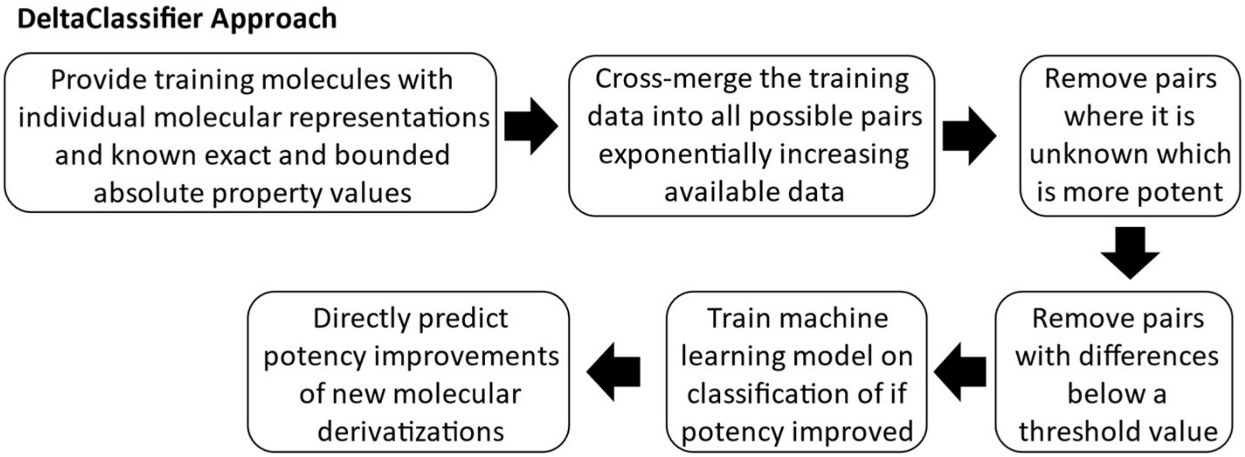
Supplementary Figure 8: Percent of Unique Scaffolds Show Limited Correlation with DΔC Improvement Over Traditional Models. Scatterplots showing correlation and Pearson's r values of average deep DeltaClassifer (DΔC) performance improvement over Random Forest (RF), XGBoost (XGB), and Chemprop (CP) following 3x10 cross-validation for 230 ChEMBL datasets with the percent of unique Murcko scaffolds within each dataset in terms of accuracy, F1 score, and AUC.



Supplementary Figure 9: Percent of Unique Scaffolds Show Limited Correlation with ΔCL Improvement Over Traditional Models. Scatterplots showing correlation and Pearson's r values of average tree-based DeltaClassifer (ΔCL) performance improvement over Random Forest (RF), XGBoost (XGB), and Chemprop (CP) following 3x10 cross-validation for 230 ChEMBL datasets with the percent of unique Murcko scaffolds within each dataset in terms of accuracy, F1 score, and AUC.

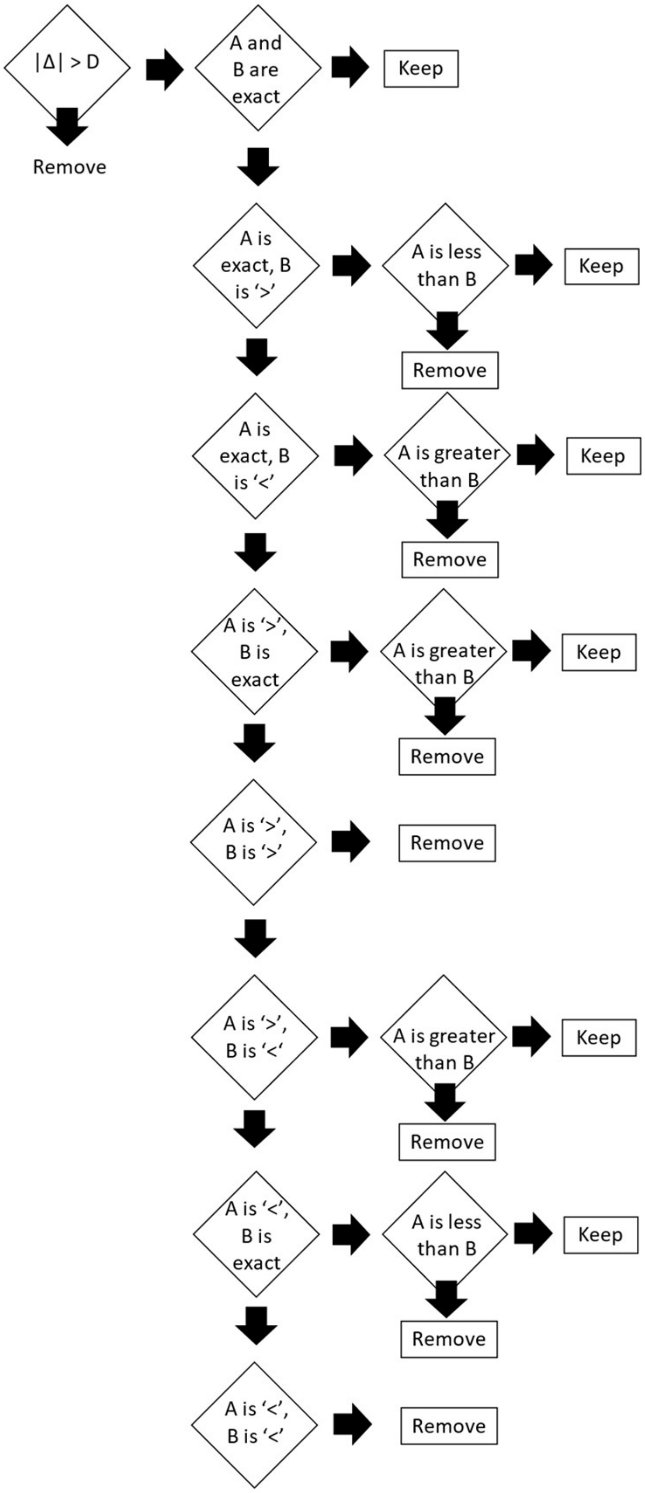

Standard Approach
Provide training molecules with individual molecular representations and known exact absolute property values
Train machine learning model on absolute property values
Predict absolute property values for new molecules
Manually subtract predicted values to approximate property changes
Classify molecular improvements based on this change

Supplementary Figure 10: Standard Regression Approach to Classify Potency Improvements.


DeltaClassifier Approach
Provide training molecules with individual molecular representations and known exact and bounded absolute property values
Cross-merge the training data into all possible pairs exponentially increasing available data
Remove pairs where it is unknown which is more potent
Remove pairs with differences below a threshold value
Train machine learning model on classification of if potency improved
Directly predict potency improvements of new molecular derivatizations

Supplementary Figure 11: DeltaClassifier Approach to Classify Potency Improvements.

Demilitarized Removal of Pairs with Unknown Potency



Supplementary Figure 12: Removal of Pairs with Unknown Potency or Differences Below Demilitarization Threshold. 'A' represents the known potency value for the first molecule within the pair. **'**B' represents the known potency value for the second molecule within the pair. 'Δ' represents the difference between 'A' and 'B'. 'D' represents the threshold set for demilitarization. Rightward facing arrows indicate 'yes' to the scenario proposed within the diamond while downward facing arrows indicate 'no'.

Supplementary Tables

Supplementary Table 1: Potency Distribution of Available IC$_{50}$ Data. Percentages of exact, bounded, and all datapoints that are above or below 1 μM in potency and average number of unique scaffolds in each dataset for our 230 IC$_{50}$ datasets.

	Exact	Bounded	All
> 1 μM	63.4%	12.8%	54.4%
< 1 μM	36.6%	87.2%	45.6%
Average Unique Scaffolds	167	52	208

Supplementary Table 2: Results for 1x10-Fold Cross-Validation Tested on Demilitarized Data. Average and standard deviation of accuracy, F1 score, and AUC are presented for all models following removal of molecular pairs with differences greater than 0.1 pIC$_{50}$ in the test set across our 230 IC$_{50}$ datasets. Highest statistically significant overall performances across all models are underlined. Highest statistically significant performances within each model family (traditional models, tree-based Δ classifiers, and deep Δ classifiers) are bolded.

	Traditional Methods (Single Molecule Regression)			Tree-Based Δ Classifiers (XGBoost)			Deep Δ Classifiers (Chemprop)		
Metric	**RF**	**XGB**	**CP**	**ΔCLOE**	**ΔCLAD**	**ΔCL**	**DΔCOE**	**DΔCAD**	**DΔC**
Accuracy	**0.795 ±0.057**	0.785 ±0.061	0.748 ±0.069	0.785 ±0.059	**0.824 ±0.048**	**0.824 ±0.048**	0.797 ±0.056	**0.836 ±0.045**	**0.836 ±0.045**
F1 Score	**0.795 ±0.057**	0.785 ±0.061	0.748 ±0.069	0.783 ±0.062	**0.823 ±0.048**	**0.824 ±0.048**	0.796 ±0.059	**0.835 ±0.045**	**0.836 ±0.045**
ROCAUC	**0.874 ±0.063**	0.861 ±0.069	0.825 ±0.081	0.863 ±0.065	**0.901 ±0.048**	**0.901 ±0.047**	0.874 ±0.060	**0.910 ±0.042**	**0.910 ±0.042**

Supplementary Table 3: Results for 1x10-Fold Cross-Validation Tested on All Datapoints. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined. Highest statistically significant performances within each model family (traditional models, tree-based Δ classifiers, and deep Δ classifiers) are bolded.

	Traditional Methods (Single Molecule Regression)			**Tree-Based Δ Classifiers (XGBoost)**			**Deep Δ Classifiers (Chemprop)**		
Metric	**RF**	**XGB**	**CP**	**ΔCLOE**	**ΔCLAD**	**ΔCL**	**DΔCOE**	**DΔCAD**	**DΔC**
Accuracy	**0.785** **±0.055**	0.776 ±0.058	0.742 ±0.065	0.770 ±0.056	**0.807** **±0.048**	**0.804** **±0.048**	0.780 ±0.054	**0.817** **±0.045**	**0.815** **±0.045**
F1 Score	**0.780** **±0.057**	0.770 ±0.060	0.736 ±0.068	0.764 ±0.060	**0.803** **±0.050**	**0.801** **±0.049**	0.775 ±0.058	**0.813** **±0.047**	**0.812** **±0.047**
ROCAUC	**0.857** **±0.064**	0.845 ±0.069	0.810 ±0.080	0.848 ±0.065	**0.886** **±0.050**	**0.885** **±0.050**	0.859 ±0.060	**0.896** **±0.046**	**0.895** **±0.045**

Supplementary Table 4: Results for 1x10-Fold Cross-Validation Tested Without Same Molecule Pairs. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models following removal of molecular pairs of the same molecule in the test set across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined. Highest statistically significant performances within each model family (traditional models, tree-based Δ classifiers, and deep Δ classifiers) are bolded.

	Traditional Methods (Single Molecule Regression)			**Tree-Based Δ Classifiers (XGBoost)**			**Deep Δ Classifiers (Chemprop)**		
Metric	**RF**	**XGB**	**CP**	**ΔCLOE**	**ΔCLAD**	**ΔCL**	**DΔCOE**	**DΔCAD**	**DΔC**
Accuracy	**0.781** **±0.057**	0.771 ±0.060	0.736 ±0.067	0.772 ±0.058	**0.811** **±0.049**	**0.81** **±0.049**	0.784 ±0.055	**0.822** **±0.046**	**0.822** **±0.046**
F1 Score	**0.780** **±0.057**	0.770 ±0.060	0.736 ±0.068	0.769 ±0.061	**0.809** **±0.050**	**0.81** **±0.049**	0.782 ±0.058	**0.821** **±0.047**	**0.821** **±0.047**
ROCAUC	**0.861** **±0.064**	0.848 ±0.070	0.813 ±0.081	0.850 ±0.066	**0.889** **±0.050**	**0.889** **±0.050**	0.862 ±0.061	**0.899** **±0.045**	**0.898** **±0.045**

Supplementary Table 5: Demilitarization Parameter Optimization for 1x10-Fold Cross-Validation Tested. Average and standard deviation of rankings of z-scores for accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC$_{50}$ datasets. Highest statistically significant performances across all models are underlined and bolded.

	Deep Δ Classifiers (Chemprop)		
Metric	**0.1 pIC$_{50}$**	**0.5 pIC$_{50}$**	**1.0 pIC$_{50}$**
Accuracy	**0.815 ±0.045**	0.812 ±0.046	0.807 ±0.049
F1 Score	**0.812 ±0.047**	0.81 ±0.048	0.804 ±0.05
ROCAUC	**0.895 ±0.045**	0.893 ±0.047	0.89 ±0.051

Supplementary Table 6: Demilitarization Parameter Optimization for 1x10-Fold Cross-Validation Tested Without Same Molecule Pairs. Average and standard deviation of rankings of z-scores for accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC$_{50}$ datasets. Highest statistically significant performances across all models are underlined and bolded.

	Deep Δ Classifiers (Chemprop)		
Metric	**0.1 pIC$_{50}$**	**0.5 pIC$_{50}$**	**1.0 pIC$_{50}$**
Accuracy	**0.822 ±0.046**	0.819 ±0.047	0.814 ±0.05
F1 Score	**0.821 ±0.046**	0.819 ±0.047	0.813 ±0.05
ROCAUC	**0.898 ±0.045**	0.897 ±0.047	0.893 ±0.051

Supplementary Table 7: Y-Shuffling Adversarial Control Experiment Collapses Model Performance. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for 1x10-fold cross-validation following Y-shuffling across our 230 IC_{50} datasets. Y-shuffling destroys the correlation between input and output variables and therefore creates a more “random” model with an accuracy, F1-score, and ROC-AUC close to 0.5.

	Δ Classifiers (after Y-shuffling)	
Metric	**ΔCL**	**DΔC**
Accuracy	0.554 ±0.052	0.551 ±0.053
F1 Score	0.544 ±0.052	0.542 ±0.052
ROCAUC	0.578 ±0.074	0.577 ±0.077

Supplementary Table 8: Ranking of Model Performance for 3x10-Fold Cross-Validation Tested on Demilitarized Data. Average and standard deviation of rankings of z-scores for accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC_{50} datasets.

	Traditional Methods (Single Molecule Regression)			**Deep Δ Classifiers (Chemprop)**	
Metric	**RF**	**XGB**	**CP**	**ΔCL**	**DΔC**
Accuracy	2.907 ±0.799	3.837 ±0.595	4.835 ±0.560	2.130 ±0.724	1.291 ±0.652
F1 Score	2.913 ±0.801	3.841 ±0.593	4.830 ±0.578	2.128 ±0.716	1.287 ±0.637
ROCAUC	2.857 ±0.788	3.865 ±0.587	4.813 ±0.564	2.091 ±0.839	1.374 ±0.705

Supplementary Table 9: Performance of k-NN Approach Compared to DeltaClassifiers. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for 1x10-fold cross-validation following removal of molecular pairs with differences greater than 0.1 pIC_{50} in the test set across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined. Statistically significant improvements compared to the k-nearest neighbours algorithm (k-NN) are bolded.

	Parameter Free	Δ Classifiers	
Metric	**k-NN**	**ΔCL**	**DΔC**
Accuracy	0.781 ±0.064	**0.824 ±0.048**	**0.836 ±0.045**
F1 Score	0.780 ±0.066	**0.824 ±0.048**	**0.836 ±0.045**
ROCAUC	0.860 ±0.071	**0.901 ±0.047**	**0.910 ±0.042**

Supplementary Table 10: Results for 80-20 Scaffold Split on Demilitarized Data. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for five models across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined and bolded.

	Traditional Methods (Single Molecule Regression)				Δ Classifiers	
Metric	**RF**	**XGB**	**CP**		**ΔCL**	**DΔC**
Accuracy	0.730 ±0.074	0.689 ±0.087	0.719 ±0.074		0.742 ±0.071	**0.766 ±0.073**
F1 Score	0.730 ±0.074	0.689 ±0.087	0.718 ±0.074		0.742 ±0.071	**0.766 ±0.072**
ROCAUC	0.805 ±0.086	0.753 ±0.111	0.790 ±0.086		0.814 ±0.082	**0.839 ±0.084**

Supplementary Table 11: Results for 80-20 Scaffold Split on All Datapoints Without Demilitarization. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for five models across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined and bolded.

	Traditional Methods (Single Molecule Regression)		Δ Classifiers

Metric	RF	XGB	CP		ΔCL	DΔC
Accuracy	0.722 ±0.071	0.711 ±0.070	0.684 ±0.083		0.729 ±0.069	**0.751 ±0.071**
F1 Score	0.718 ±0.072	0.707 ±0.071	0.680 ±0.084		0.728 ±0.069	**0.750 ±0.071**
ROCAUC	0.792 ±0.085	0.778 ±0.085	0.742 ±0.107		0.802 ±0.081	**0.826 ±0.083**

Supplementary Table 12: Results for 80-20 Scaffold Split Without Demilitarization or Same Molecule Pairs. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for five models across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined and bolded.

	Traditional Methods (Single Molecule Regression)				Δ Classifiers	
Metric	**RF**	**XGB**	**CP**		**ΔCL**	**DΔC**
Accuracy	0.719 ±0.072	0.680 ±0.084	0.708 ±0.071		0.732 ±0.070	**0.754 ±0.071**
F1 Score	0.718 ±0.072	0.680 ±0.084	0.707 ±0.071		0.732 ±0.069	**0.754 ±0.071**
ROCAUC	0.793 ±0.085	0.743 ±0.108	0.779 ±0.085		0.803 ±0.082	**0.828 ±0.084**

Supplementary Table 13: Results for 1x10-Fold Cross-Validation Tested Without Same Molecule Pairs and Bounded Data. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models following removal of molecular pairs of the same molecule and molecular pairs incorporating a molecule with a bounded IC_{50} value in the test set across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined. Highest statistically significant performances within each model family (traditional models, tree-based Δ classifiers, and deep Δ classifiers) are bolded.

	Traditional Methods (Single Molecule Regression)			Tree-Based Δ Classifiers (XGBoost)			Deep Δ Classifiers (Chemprop)		
Metric	RF	XGB	CP	ΔCLOE	ΔCLAD	ΔCL	DΔCOE	DΔCAD	DΔC
Accuracy	**0.791 ±0.047**	0.784 ±0.049	0.747 ±0.052	**0.784 ±0.048**	0.779 ±0.052	0.779 ±0.052	**0.792 ±0.048**	**0.790 ±0.049**	**0.791 ±0.049**
F1 Score	**0.790 ±0.048**	0.782 ±0.050	0.746 ±0.053	**0.781 ±0.050**	0.777 ±0.054	0.778 ±0.053	**0.790 ±0.050**	**0.788 ±0.052**	**0.790 ±0.051**
ROCAUC	**0.872 ±0.049**	0.863 ±0.052	0.827 ±0.062	**0.863 ±0.052**	0.857 ±0.058	0.858 ±0.057	**0.871 ±0.051**	0.867 ±0.053	0.867 ±0.052

Supplementary Table 14: Results for 1x10-Fold Cross-Validation Tested on Demilitarized Non-Matching Scaffold Pairs. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined. Highest statistically significant performances within each model family (traditional models, tree-based Δ classifiers, and deep Δ classifiers) are bolded.

	Traditional Methods (Single Molecule Regression)			Tree-Based Δ Classifiers (XGBoost)			Deep Δ Classifiers (Chemprop)		
Metric	RF	XGB	CP	ΔCLOE	ΔCLAD	ΔCL	DΔCOE	DΔCAD	DΔC
Accuracy	**0.797 ±0.058**	0.787 ±0.062	0.751 ±0.070	0.787 ±0.060	**0.826 ±0.049**	**0.827 ±0.049**	0.800 ±0.057	**0.838 ±0.045**	**0.839 ±0.045**
F1 Score	**0.797 ±0.058**	0.787 ±0.062	0.751 ±0.070	0.786 ±0.062	**0.826 ±0.049**	**0.827 ±0.049**	0.798 ±0.059	**0.838 ±0.046**	**0.838 ±0.045**
ROCAUC	**0.875 ±0.063**	0.863 ±0.069	0.828 ±0.082	0.865 ±0.065	**0.903 ±0.048**	**0.902 ±0.047**	0.876 ±0.060	**0.912 ±0.042**	**0.912 ±0.042**

Supplementary Table 15: Ranking of Model Performance for 1x10-Fold Cross-Validation Tested on Demilitarized Data for Non-Matching Scaffold Pairs. Average and standard deviation of rankings of z-scores for accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC_{50} datasets.

	Traditional Methods	Δ Classifiers

	(Single Molecule Regression)				(XGBoost and Chemprop)	
Metric	RF	XGB	CP		ΔCL	DΔC
Accuracy	2.961 ±0.820	3.778 ±0.709	4.796 ±0.596		2.089 ±0.766	1.376 ±0.766
F1 Score	2.941 ±0.834	3.774 ±0.711	4.8 ±0.594		2.098 ±0.785	1.387 ±0.761
ROCAUC	2.863 ±0.830	3.82 ±0.702	4.785 ±0.593		2.133 ±0.839	1.400 ±0.761

Supplementary Table 16: Results for 1x10-Fold Cross-Validation Tested on Demilitarized Matching Scaffold Pairs. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC_{50} datasets. Highest statistically significant overall performances across all models are underlined. Highest statistically significant performances within each model family (traditional models, tree-based Δ classifiers, and deep Δ classifiers) are bolded.

	Traditional Methods (Single Molecule Regression)				Tree-Based Δ Classifiers (XGBoost)				Deep Δ Classifiers (Chemprop)		
Metric	RF	XGB	CP		ΔCLOE	ΔCLAD	ΔCL		DΔCOE	DΔCAD	DΔC
Accuracy	**0.670 ±0.099**	**0.670 ±0.093**	0.601 ±0.097		0.657 ±0.087	**0.674 ±0.084**	**0.675 ±0.080**		0.678 ±0.092	**0.701 ±0.085**	**0.699 ±0.083**
F1 Score	**0.668 ±0.101**	**0.667 ±0.095**	0.601 ±0.097		0.635 ±0.107	0.660 ±0.092	**0.676 ±0.079**		0.669 ±0.102	0.695 ±0.088	**0.699 ±0.083**
ROCAUC	**0.733 ±0.114**	0.723 ±0.116	0.638 ±0.124		0.720 ±0.108	**0.744 ±0.100**	**0.744 ±0.099**		0.740 ±0.114	**0.768 ±0.099**	**0.767 ±0.098**

Supplementary Table 17: Ranking of Model Performance for 1x10-Fold Cross-Validation Tested on Demilitarized Data for Matching Scaffold Pairs. Average and standard deviation of rankings of z-scores for accuracy, F1 score, and ROCAUC are presented for all models across our 230 IC_{50} datasets.

	Traditional Methods (Single Molecule Regression)		Deep Δ Classifiers (Chemprop)

Metric	RF	XGB	CP	ΔCL	DΔC
Accuracy	2.748 ±1.149	2.941 ±1.222	4.398 ±1.046	2.989 ±1.183	1.924 ±1.190
F1 Score	2.772 ±1.145	2.983 ±1.236	4.370 ±1.078	2.952 ±1.192	1.924 ±1.208
ROCAUC	2.757 ±1.122	3.100 ±1.173	4.498 ±0.981	2.759 ±1.216	1.887 ±1.155

Supplementary Table 18: Model Performance for Enzyme Class 1. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models across our 28 IC_{50} datasets for targets in enzyme class 1. Statistically significant performances over traditional methods are bolded.

	Traditional Methods **(Single Molecule Regression)**			**Deep Δ Classifiers** **(Chemprop)**	
Metric	**RF**	**XGB**	**CP**	**ΔCL**	**DΔC**
Accuracy	0.792 ±0.066	0.783 ±0.071	0.747 ±0.069	**0.822** **±0.049**	**0.837** **±0.039**
F1 Score	0.792 ±0.066	0.783 ±0.071	0.747 ±0.069	**0.823** **±0.049**	**0.837** **±0.039**
ROCAUC	0.869 ±0.071	0.857 ±0.080	0.821 ±0.083	**0.899** **±0.047**	**0.910** **±0.035**

Supplementary Table 19: Model Performance for Enzyme Class 2. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models across our 69 IC_{50} datasets for targets in enzyme class 2. Statistically significant performances over traditional methods are bolded.

	Traditional Methods **(Single Molecule Regression)**			**Deep Δ Classifiers** **(Chemprop)**	
Metric	**RF**	**XGB**	**CP**	**ΔCL**	**DΔC**
Accuracy	0.795 ±0.051	0.786 ±0.052	0.749 ±0.065	**0.824** **±0.047**	**0.835** **±0.046**
F1 Score	0.795 ±0.051	0.786 ±0.052	0.749 ±0.065	**0.824** **±0.047**	**0.834** **±0.046**
ROCAUC	0.873	0.862	0.826	**0.900**	**0.908**

	±0.055	±0.056	±0.077		**±0.044**	**±0.042**

Supplementary Table 20: Model Performance for Enzyme Class 3. Average and standard deviation of accuracy, F1 score, and ROCAUC are presented for all models across our 73 IC_{50} datasets for targets in enzyme class 3. Statistically significant performances over traditional methods are bolded.

	Traditional Methods (Single Molecule Regression)				**Deep Δ Classifiers (Chemprop)**	
Metric	**RF**	**XGB**	**CP**		**ΔCL**	**DΔC**
Accuracy	0.797 ±0.062	0.786 ±0.067	0.749 ±0.074		**0.829 ±0.043**	**0.841 ±0.042**
F1 Score	0.797 ±0.062	0.785 ±0.067	0.749 ±0.074		**0.829 ±0.043**	**0.841 ±0.042**
ROCAUC	0.876 ±0.072	0.862 ±0.079	0.825 ±0.088		**0.906 ±0.042**	**0.915 ±0.040**

Supplementary Table 21: Model Performance for Enzyme Class 4. Accuracy, F1 score, and ROCAUC are presented for all models across our 1 IC_{50} dataset for targets in enzyme class 4.

	Traditional Methods (Single Molecule Regression)				**Deep Δ Classifiers (Chemprop)**	
Metric	**RF**	**XGB**	**CP**		**ΔCL**	**DΔC**
Accuracy	0.808	0.798	0.788		0.871	0.879
F1 Score	0.808	0.798	0.788		0.872	0.879
ROCAUC	0.900	0.891	0.881		0.941	0.948

Supplementary Table 22: Model Performance for Enzyme Class 5. Accuracy, F1 score, and ROCAUC are presented for all models across our 3 IC_{50} dataset2 for targets in enzyme class 5.

	Traditional Methods (Single Molecule Regression)				**Deep Δ Classifiers (Chemprop)**	
Metric	**RF**	**XGB**	**CP**		**ΔCL**	**DΔC**
Accuracy	0.781	0.777	0.788		0.860	0.866

	±0.059	±0.052	±0.030		±0.029	±0.017
F1 Score	0.781 ±0.059	0.777 ±0.052	0.788 ±0.030		0.860 ±0.029	0.866 ±0.017
ROCAUC	0.864 ±0.071	0.858 ±0.070	0.871 ±0.038		0.931 ±0.022	0.939 ±0.012

Supplementary Table 23: Related Approaches to Compare Properties of Molecular Pairs.

Approach	Predictive Target	Citation
Siamese neural network	Molecular similarity	M. K. Altalib and N. Salim, *ACS Omega*, 2022, **7**, 4769–4786.
Siamese neural network	Bioactivity	D. Fernández-Llaneza, S. Ulander, D. Gogishvili, E. Nittinger, H. Zhao and C. Tyrchan, *ACS Omega*, 2021, **6**, 11086–11094.
Siamese neural network	Toxicity	H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, *ACS Cent Sci*, 2017, **3**, 283–293.
Siamese neural network	Drug-drug interactions	K. Schwarz, A. Allam, N. A. Perez Gonzalez and M. Krauthammer, *BMC Bioinformatics*, 2021, **22**, 1–19.
Siamese neural network	Relative free energy of binding	A. T. McNutt and D. R. Koes, *J Chem Inf Model*, 2022, **62**, 1819–1829.
Siamese neural network	Transcriptional response similarity	M. Jeon, D. Park, J. Lee, H. Jeon, M. Ko, S. Kim, Y. Choi, A.-C. Tan and J. Kang, *Bioinformatics*, 2019, **35**, 5249–5256.
Siamese Neural Network	ADMET properties	Y. Zhang, J. Menke, J. He, E. Nittinger, C. Tyrchan, O. Koch and H. Zhao, *J Cheminform*, 2023, **15**, 75.
Kernel-based ranking algorithms	Potency	S. Agarwal, D. Dugar and S. Sengupta, *J Chem Inf Model*, 2010, **50**, 716–731.
Learning-to-rank framework	Potency	K. L. Saar, W. McCorkindale, D. Fearon, M. Boby, H. Barr, A. Ben-Shmuel, C. M. Consortium, N. London, F. von Delft and J. D. Chodera, *Proceedings of the National Academy of Sciences*, 2023, **120**, e2214168120.
Learning-to-rank framework	Potency	A. Morris, W. McCorkindale, N. Drayman, J. D. Chodera, S. Tay, N. London and Covid Moonshot Consortium, *Chemical Communications*, 2021, **57**, 5909–5912.
QSAR modeling	Potency	K. Matsumoto, T. Miyao and K. Funatsu, *ACS Omega*, 2021, **6**, 11964–11973.